INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *Project-Team LEAR*

# *Learning and Recognition in Vision*

*Rhône-Alpes*

THEME 3B

# Contents

# 1 Team

LEAR *is part of the GRAVIR/IMAG laboratory, a Joint Research Unit of INRIA, the Centre National de Recherche Scientifique (CNRS), the Institut National Polytechnique de Grenoble (INPG) and the Université Joseph Fourier (UJF).* LEAR *was created officially on 1st July 2003, but this report summarizes the activities of its members throughout all of 2003.*

**Head of project team**

Cordelia Schmid [CR, INRIA]

**Vice-head of project team**

Bill Triggs [CR, CNRS]

**Administrative assistant**

Anne Pasteur

**Faculty member**

Roger Mohr [Professor, ENSIMAG]

**Technical staff**

Michael Sdika [09/2003-09/2004]

**Post-doctoral fellow**

Jianguo Zhang [INRIA scholarship, 12/2003-11/2004]

**PhD students**

Ankur Agarwal [INRIA scholarship]

Charles Bouveyron [MENESR scholarship, starting 09/2003]

Guillaume Bouchard [INRIA scholarship]

Navneet Dalal [INRIA scholarship]

Gyuri Dorkó [INRIA scholarship]

**Student interns**

Peter Carbonetto [master, 10/2003-06/2004]

Christine Dratva [DEA IVR, 04/2003-08/2003]

Radek Filip [DEA IVR, 04/2003-06/2003]

Remy Bernard [DEA IVR, 04/2003-06/2003]

Salil Jain [DEA IVR, 06/2003-08/2005, French Embassy scholarship]

Shakti Kamal [third year IIT Delhi, 06/2003-11/2003]

Nipun Kwatra [third year IIT Delhi, 05/2003-07/2003, INRIA-IIT scholarship]

**Visiting scientist**

Frédéric Jurie [CR, CNRS, 09/2003-08/2004]

**Visiting engineer**

Marius Malciu [Pandora Studio, 06/2003-08/2003]

# 2 Overall objectives

LEAR's main focus of research is learning based approaches to visual object recognition and scene interpretation, particularly for image retrieval and video indexing. Understanding the content of everyday images and videos is one of the most challenging problems in computer vision. The extent to which we can do this is currently limited, but we believe that very significant advances will be made over the next few years by combining emerging statistical learning techniques with state of the art image descriptors. This field is also close to a major threshold of applicability: even partial solutions are likely to enable many new applications.

LEAR's main research areas are:

- **Image description.** Many efficient lighting and viewpoint invariant image descriptors are now available, such as for example affine-invariant interest points. Our current research aims to extend these techniques to describe textures, to define more powerful similarity and saliency measures and to characterize 2D and 3D shape information.
- **Learning.** Our research on machine learning and statistical modelling is mainly aimed at improving their applicability to visual recognition and computer vision. It includes both the selection, evaluation and adaptation of existing methods, and the development of new ones designed to take vision specific constraints into account. Particular challenges include: dealing with the huge amounts of data that image and video collections contain; handling large rich natural class hierarchies rather than just simple yes/no classifiers; and capturing enough information about the domain to allow generalization from just a few images, rather than from large carefully marked-up training databases.

- **Recognition.** Visual object recognition requires the construction of exploitable visual models for both particular objects and object categories. Achieving good invariance to viewpoint, lighting, occlusion and background is challenging even for exactly known rigid objects, and these difficulties are greatly compounded when reliable generalization across object categories is needed. Our research combines advanced image description techniques with learning for good invariance and generalization. Currently the selection and coupling of image descriptors and learning techniques is done by hand, and one significant challenge is the automation of this process, for example using automatic feature selection and statistically-based validation diagnostics.

# 3 Scientific foundations

## 3.1 Image description

We believe that the extraction of robust image descriptors is a critical component of any visual recognition system, and even though many efficient descriptors are already available, further research is clearly needed in this area. One can go a certain distance using simplistic descriptors, but their unreliability and lack of invariance puts a heavy burden on the learning method and the training data and ultimately limits the performance that can be achieved. Better descriptors allow simpler learning methods to be used and produce better separation of classes, potentially allowing generalization from just a few examples instead of requiring large, carefully engineered training databases.

The kinds of descriptors that we advocate have a certain number of basic properties:

- **Locality and redundancy:** For resistance to changes of background and occlusions, reduced sensitivity to changes of viewpoint and variable intra-class geometry, and robustness against individual feature extraction failures, descriptors should have relatively small spatial support, but there should be many of them in each image. Schemes based on collections of image patches or fragments are more robust and better adapted to object-level queries than global whole-image descriptors.

- **Salience:** Fragments are not very useful unless they can be extracted automatically and found again in other images. Hence, rather than using general fragments, we focus on local descriptors based at particularly salient points — "keypoints" or "points of interest". This gives a sparser and hence more efficient representation, and one that can be constructed automatically in a preprocessing step. To be useful, such points must be accurately relocalizable in other images, with respect to both position and scale.

- **Photometric and geometric invariance:** The interest points and their descriptors should have an appropriate degree of invariance to changes of illumination and variations of local image geometry induced by changes of viewpoint, viewing distance, and local intra-class variability. In practice, geometric invariance is usually approximated by invariance to Euclidean, similarity or affine transforms of the local image.

- **Informativeness:** Notwithstanding all of the above types of invariance, the descriptors should be *informative* in the sense that they are rich sources of information about image content that can

easily be exploited in scene characterization and object recognition tasks. Images contain a lot of variety, so this requires relatively high dimensionality. Just as importantly, the useful information should be manifest, not hidden in obscure high-order correlations between coefficients. Image formation is essentially a spatial process, so in practice this favours descriptors that code relative position information manifestly (e.g. context-style descriptors rather than moments or Fourier descriptors).

Our current research in this area is focused on creating detectors and descriptors that are better adapted to particular kinds of imagery, incorporating spatial neighbourhood and region constraints to improve informativeness, and extending the scheme to cover different kinds of locality.

### 3.2 Learning

We are interested in learning and statistics mainly as technologies for attacking difficult vision problems, so we take an eclectic approach, using a wide variety of techniques ranging from classical statistical generative and discriminative models to modern kernel, margin and boosting based machines. Parameter-rich models and limited training data are the norm in vision, so overfitting needs to be controlled by various types of regularization, model and feature selection, and dimensionality reduction methods, after being measured using methods such as cross-validation, information criteria and capacity bounds. Visual descriptors tend to be high dimensional and they typically contain some redundancy, so we often preprocess data using techniques such as PCA and its nonlinear variants, ICA, and LLE/Isomap, to reduce it to a more manageable dimensionality. To capture the shapes of complex probability distributions over high dimensional descriptor spaces, we either fit mixture models and similar structured semi-parametric probability models, or reduce them to histograms using vector quantization techniques such as K-means. Missing data is common owing to unknown class labels, feature detection failures, occlusions and intra-class variability, so we often need to use completion techniques such as Expectation Maximization. On the discriminative side, machine learning techniques such as Support Vector Machines, Relevance Vector Machines, and Boosting, are used to produce flexible classifiers and regression methods based on visual descriptors. Visual categories have a rich nested structure, so techniques that handle large numbers of classes and nested classes are especially interesting to us. Images contain huge amounts of data, so we need to use algorithms suited to large-scale learning problems, and it is expensive and tedious to label large numbers of training images, so unsupervised, semi-supervised and transductive learning methods are of particular interest. Weakly labelled data is also common — for example one may be told that a training image contains an object of some class, but not where the object is in the image — and variants of unsupervised, correlational, and co- learning are useful for handling this.

We keep up to date on learning technology by maintaining active links with both the statistics community, most notably via collaborations with the INRIA projects MISTIS and SELECT (formerly IS2), and the machine learning one, most notably via the EU project LAVA and the Network of Excellence PASCAL.

### 3.3 Recognition

The current state of progress in visual recognition shows clearly that combining advanced image descriptors with modern learning and statistical modelling techniques has the potential to produce very

significant advances. We believe that taken together and tightly integrated, these techniques have the potential to to make visual recognition a widespread technology.

The kind of process that we advocate makes full use of the unusual robustness and richness of our image description methods (see §3.1) to provide a vocabulary of base features that already goes a long way towards characterizing the category being recognized. The final learning based classifier is thus mainly responsible for extending the model to larger amounts of intra-class variation and gross changes of aspect or viewpoint, and for capturing the subtler higher order correlations that are needed to fine tune the base performance. That said, our approach is not simply feature extraction *then* learning: the integration is actually much tighter than this. Nearly every stage of our descriptor chain uses learning and statistical modelling in a fundamental way, to generate or select robust invariant features, to squeeze out redundancy and bring out informativeness. Similarly, to maximize their performance, the final learning methods use descriptor comparison metrics (kernels, reference densities, structural models) that are intimately based on the statistical properties and invariances (or lack thereof) of the learned descriptors.

# 4 Application domains

A solution to the general object recognition problem will enable a wide range of applications including defense, health care, human-computer interaction, image retrieval and data mining, industrial and personal robotics, manufacturing, scientific image analysis, space exploration, surveillance and security, and transportation. In fact, with the ever expanding array of image sources, some form of automatic object recognition technology must eventually be an integral part of every information system. Even partial solutions are likely to enable many applications.

Our project's main application domain is image and video indexing. This is an area with huge potential. For example, it is estimated that 96% of all data currently generated by humanity is personal images and home videos[1]. Currently, we are working on developing indexing techniques for camera equipped handheld devices such as personal digital assistants, on object-level structuring and indexing of feature film videos, and on applying our techniques to surveillance in the context of military applications.

**A personal visual assistant** is a portable device equipped with a camera that can identify the category or instance of an object that it sees, and supply the user with associated information. A software prototype is being developed within the European project LAVA to test and validate our algorithms.

**Object-level video structuring** organizes the content of a video in terms of the objects and actions in it, and thus allows the user to browse and access the video in terms of semantically meaningful units. Given a set of actors, scenes or actions, the method can find other locations in the video where that combination of entities occurred. The software for these tools is being developed within the European project VIBES.

**Surveillance** requires the detection and recognition of objects. In our case, a military application, the camera is static and the detection is therefore relatively straightforward. The subsequent recognition should then differentiate between different types of vehicles.

[1] `http://www.sims.berkeley.edu/research/projects/how-much-info/summary.html`

# 5 Software

## 5.1 Software for computing local invariant features

**Contributed by**: Cordelia Schmid, Michael Sdika, Gyuri Dorko, Krystian Mikolajczyk [Oxford].

The local feature extraction programs developed during the PhD of K. Mikolajczyk [Mik02] have been improved, and executables are available on our web page `http://www.inrialpes.fr/lear/downloads.html`. The different detectors and descriptors were compared in our evaluation of interest points and regions [7]. The images used in this evaluation are also available on the web at `http://www.inrialpes.fr/movi/people/Mikolajczyk/Database`.

# 6 New results

## 6.1 Image description

**Contributed by**: Cordelia Schmid, Michael Sdika, Frédéric Jurie, Bill Triggs, Remy Bernard, Krystian Mikolajczyk [Oxford], Andrew Zisserman [Oxford], Fred Rothganger [UIUC], Jean Ponce [UIUC].

**Keywords**: photometric invariants, grey-level descriptors, shape features, performance evaluation.

### 6.1.1 Affine-invariant descriptors

We have developed scale- and affine-invariant salient point detectors [5, 7] that give excellent performance for recognizing both specific objects and scenes, and texture and object classes [6, 8].

Scale invariance is obtained by searching for maxima in scale-space. Different functions can be used to construct the scale-space, for example the Laplacian and the Hessian. A combination of Harris interest points computed in scale-space with a scale selection based on the Laplacian has shown very good performance. However, in the presence of significant viewpoint changes, scale invariance alone no longer suffices for reliable recognition, and an extension to affine invariance is necessary. This is obtained by running an iterative affine warping procedure that reduces the interest point's second-moment matrix to normal form. For each point we then obtain an associated affine-invariant region on which a conventional descriptor can be computed (see figure 1). A performance evaluation has shown that the points and their regions can be detected repeatedly in the presence of significant scale changes (up to a factor 4) and affine deformations (viewing angle changes of up to 70 degrees).

Various other approaches for detecting affine-invariant interest points or regions have been developed at Leuven, Oxford and Prague universities, the Leuven detectors combining points and edges as well as extracting intensity maxima, the Prague one extracting maximally stable connected components. We are currently collaborating with Leuven, Oxford and Prague on a comparison of these

[Mik02] K. MIKOLAJCZYK, *Detection of local features invariant to affine transformations*, PdD Thesis, INPG, July 2002.

Figure 1: Our affine-invariant regions recover the same photometric information despite large changes of viewpoint, and hence allow rich viewpoint-invariant descriptors to be calculated. Here the method is used for wide-baseline image matching. Only the regions matched between the two images are displayed.

approaches. The preliminary results show that the different detectors yield complementary information. This comparison will be published, and we will also make the executables, the test images and the evaluation procedure available on the web.

### 6.1.2 Performance evaluation of local descriptors

Given a set of stably-detected local image regions, we can calculate local image descriptors based on them and use these for matching and recognition. The descriptors should be distinctive and at the same time robust to both changes in the illumination and viewing conditions and inaccuracies of the region detector. Many different descriptors have been proposed in the literature, but it is currently unclear which are the most appropriate for particular problems and how their performance depends on the detector. To help to clarify this, we have evaluated the pairing of several different descriptors with several different interest point detectors [20].

Our evaluation was carried out for different image types of transformations, using detection/dropout rates as the main quality criterion. By varying the value of the similarity threshold for declaring a match between two descriptors, we generate ROC (receiver operating characteristic) curves of the trade-off between the average detection rate for query images and the false positive rate across the test database (which in our case contains 1000 images and about 300 000 interest points).

We compared SIFT descriptors, steerable filters, differential invariants, complex filters, moment invariants and image cross-correlation for various different types of interest points. We find that the relative ranking of the descriptors does not depend on the underlying point detector used, and that SIFT descriptors uniformly perform best. Their success can be explained by their robustness against localization errors and small geometric distortions. Steerable filters come second, and can be considered a good choice if a lower-dimensional descriptor is needed.

Some typical results for a significant viewpoint change between the query and database image are shown in figure 2, using the affine-invariant Harris-Laplace detector. The SIFT descriptors are clearly the most robust, but note that the differences between descriptors are less important than the improve-

ment produced by enforcing affine invariance: SIFT descriptors computed with only *scale*-invariant Harris-Laplace regions ('HL SIFT' in the figure) perform worse than any of the affine descriptors shown here, as the underlying regions are not invariant enough to be redetected reliably. Steerable filters come second, but they perform significantly worse than SIFT descriptors here.

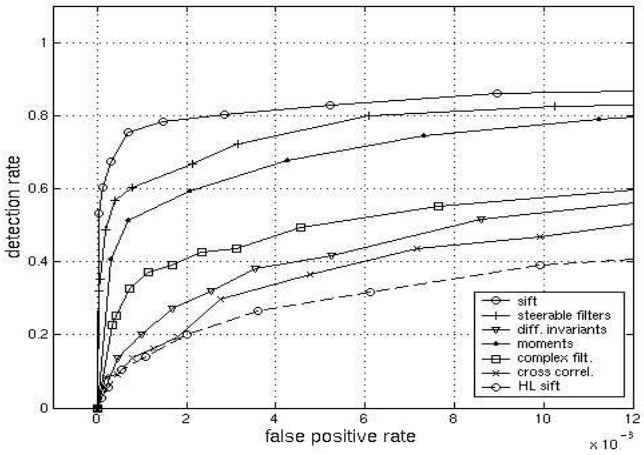


Figure 2: Descriptor evaluation for a camera viewpoint change of 60°, using affine-invariant Harris regions. HL sift is the SIFT descriptor computed for scale-invariant Harris regions.

We are working on improving the current descriptors and have so far developed a descriptor based on spin images, a rotationally invariant version of the SIFT descriptor, and extended SIFT to higher order derivatives. A comparison with existing approaches will evaluate the increase in performance.

### 6.1.3 Shape features

The descriptors presented in the previous two sections are based on the grey-level image information. Local invariant features based on such information have proven to be very successful for matching and recognition of specific textured objects. Unfortunately, for many objects the only reliable recognition cues are edges or shape, and texture cannot be used as the primary descriptor. In particular, for category-level recognition, edge and shape are often the only reliable common features between different instances of the category. To cover this case, we have recently developed two different types of scale-invariant edge descriptors. The first characterizes the type of edge pixels by the edge information in their neighbourhoods [21]. The second extracts local shape structures.

*Edge features* are edge pixels around which a scale-invariant region is estimated based on the Laplacian of the grey-level image. Each edge region is described by the distribution of relative positions and orientations of its surrounding edges. Partial descriptors allow object-based features to be obtained, even if the feature is on an object boundary.

In contrast, *local shape features* do not centre the region on an edge pixel, but instead extract influence regions that capture the local structure (shape) of the contour image. Our approach detects local shape convexities by local scale-space maximization of a robust concentricity measure, the entropy of radial gradient orientations in an annulus whose radius is defined by the scale. This is robust to clutter inside the annulus, occlusions, and spurious edge detections.

For both types of descriptors, the object is recognized in new images by a series of steps that apply progressively tighter geometric restrictions. Figure 3 shows some results on bicycles, a largely "transparent" object category that is difficult to handle by conventional methods. In this example, the object model is learnt from a single image and is correctly detected in the presence of significant scale changes. The left column shows the results for edge features and the right column for local shape features.

Figure 3: Detection of object categories based on shape features. Left column: edge features, right column: local shape features. Top row: features extracted for the training image, bottom row: detection results.

### 6.1.4 Multi-view description

It is well-established that a set of images of a rigid 3D object can be used for object recognition. However, purely image based representations are not optimal as a great deal of redundant information must be stored and they do not provide a 3D model that can be used for verification. In this work, we build a 3D model from the images and use it for recognition [23]. We use the affine-invariant patches introduced in section 6.1.1 to represent local surface appearance, and select promising matches between pairs of images or an object model and an image. We then use the geometric multi-view consistency constraints studied in the structure-from-motion literature to represent the global object structure, retain correct matches, and discard incorrect ones. Our experiments show that rigid object models can be acquired automatically from a few images (figure 4), and then used effectively for recognition tasks (figure 5).

Figure 4: Model gallery: sample input images and renderings of the corresponding models.

Figure 5: Object recognition experiments. The top row shows the model patches that were matched to the images, and the bottom row shows the models recognized, rendered in their estimated poses. Note that the teddy bear in the leftmost column is in a pose radically different from those used to acquire its model, and that there is a significant amount of clutter and occlusion in each image.

## 6.2 Learning

**Contributed by**: Bill Triggs, Guillaume Bouchard, Cordelia Schmid, Charles Bouveyron, Peter Carbonetto, Nipun Kwatra.

**Keywords**: Discriminative-generative learning, Gaussian mixture models, Support Vector Machines, semi-supervised learning, data reduction.

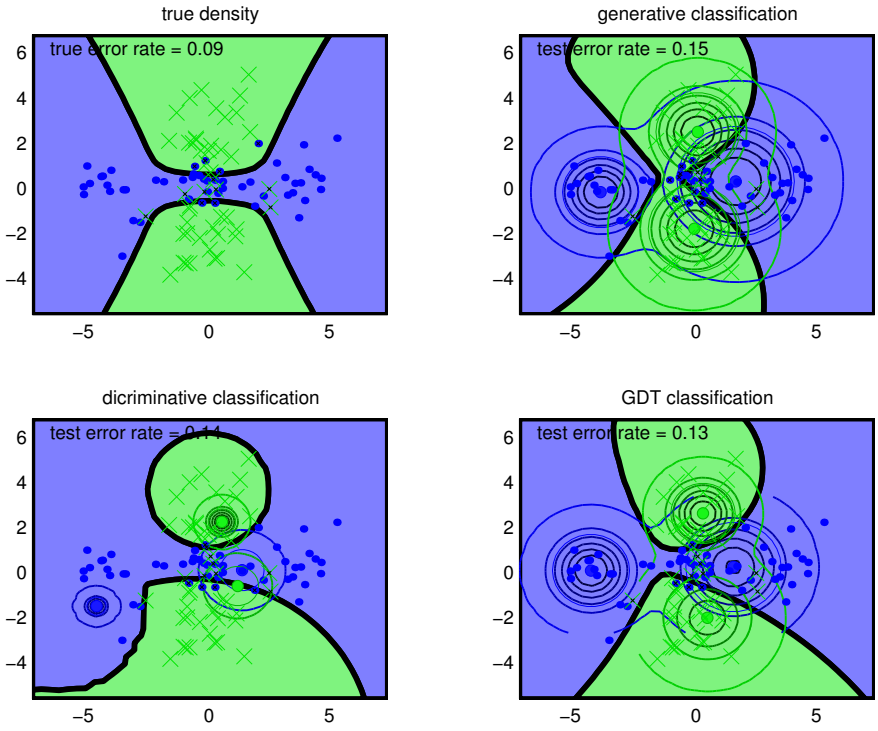


Figure 6: A simple example of combining generative and discriminative learning. Gaussian mixtures are used to fit a sample from a 2D non-Gaussian-mixture density (blue dots versus green crosses). Discriminative learning overfits, whereas in generative learning each model is unaware of the opposite class and hence fails to capture details necessary for discrimination. Combining the two gives better results than either.

### 6.2.1 Discriminative versus generative learning

There are two main statistical approaches to classification. The *discriminative* approach tries to learn a single model that directly predicts the class label from the observed input data, whereas the *generative* approach learns separate models for each class, then chooses the model (class) that best fits the observed data. Generative models are more general, they extend more naturally to complex problems (missing data, multiple classes...), and they are often simpler to learn and stabler as the classes do not interact. But if overfitting can be avoided, discriminative models typically have better classification performance: they are optimized directly for this, they have no need to model details that are important for class description but irrelevant for inter-class discrimination, and in particular they are often remarkably insensitive to certain kinds of mismodelling of the classes. In this work-in-progress, we have tried to capture some of the advantages of both approaches by combining them. Technically, this is done by fitting a discriminative model (i.e. optimizing classification performance over all classes together) that includes a generative-model based penalty term to enforce some of the regularization (but also some of the mismodelling) implicit in the full generative model. The strength of the penalty is set by cross-validation. The resulting combined model often has higher performance than either of

its parent models. See fig. 6 for a toy example.

### 6.2.2 Transductive learning for scene classification

As part of an effort to evaluate the applicability of modern semi-supervised machine learning techniques to vision problems, we studied the use of transductive support vector machines to reduce the need for supervised labelling in a simple scene classification task. We took training and test sets of images of 16 different scene classes (beaches, kitchens, rural, mountains...), calculated a global (but locally-based) descriptor vector for each image, and ran various classical and modern multiclass classifiers on the resulting descriptors, including componentwise naive Bayes, Gaussian mixtures, support vector machines with several types of kernel, and transductive SVM's with these kernels and various proportions of additional unlabelled training images. Each image was described by the distribution of its SIFT descriptor responses over affine interest points (c.f. §6.1.2). To characterize these distributions, we reduced the SIFT descriptors to 35D using global PCA for speed and stability, clustered using K-means, and quantized and counted to get each image's histogram. The overall conclusion of the study is that — at least with these descriptors and this (very small) training database — the kernel techniques perform better than the classical statistical models, and transduction does manage to reduce classification errors, but only by a modest amount, and only if most of the data is labelled. In fact, transduction actually becomes very unreliable if more than about half of the data is unlabelled. Hence, by itself, transductive learning is certainly not enough to provide human-style generalization from just one or a few images in these tests.

## 6.3 Recognition

**Contributed by**: Cordelia Schmid, Gyuri Dorko, Bill Triggs, Ankur Agarwal, Roger Mohr, Shakti Kamal, Navneet Dalal, Salil Jain, Jianguo Zhang, Svetlana Lazebnik [UIUC], Jean Ponce [UIUC], Krystian Mikolajczyk [Oxford], Andrew Zisserman [Oxford].

**Keywords**: visual models, object class recognition and detection, humans detection.

### 6.3.1 Texture recognition

In our past work, we developed a method for weakly supervised learning of visual models that can handle both complex natural textures — for example "textured" animals such as zebras and leopards — and highly structured patterns such as parts of a face [9]. The visual model is based on a two-layer image description: a set of underlying "generic" descriptors, and a learned distribution over neighbourhoods. This description method is rotationally invariant, robust to model deformations, and it efficiently characterizes "appearance-based" visual structure. The learning method is based on selecting distinctive clusters in descriptor space — descriptors common in the positive and rare in the negative examples — and using these as features for object recognition. Experimental results for "textured" animals and faces show a very good performance for retrieval as well as localization.

This representation has now been extended to allow recognition of texture patterns despite appearance variations due to non-rigid transformations and changes in viewpoint [17, 18, 19]. At the feature extraction stage, a set of affine-invariant local patches is extracted from the image. Affine-invariant

patches provide both a good level of spatial selectivity, and a texture representation that is invariant to any geometric transformation that can be locally approximated by the affine model. Each patch is characterized using an gray-level descriptor. In each image, the distribution of descriptors is summarized as a set of weighted clusters. These signatures are then compared using the Earth Mover's Distance (EMD) — a convenient and effective dissimilarity measure. For our application, the signature/EMD framework offers several important advantages. Signatures are more descriptive than histograms with fixed bins and do not require global clustering of the descriptors from all images. EMD can match signatures of different sizes, and it is not very sensitive to the number of clusters. This texture representation has been evaluated for retrieval and classification tasks using a collection of images of textured surfaces taken from different viewpoints. Figure 7 shows two sample images from each texture class. Significant viewpoint and scale changes occur within each class, and several of the classes include additional sources of variability: inhomogeneities in the texture patterns, non-rigid transformations, illumination changes, and unmodeled viewpoint-dependent appearance changes. Retrieval and classification results for these textures are excellent [19].

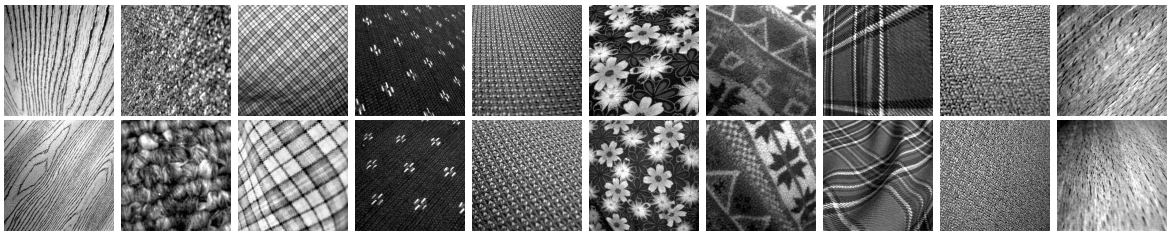

Figure 7: Samples of the ten texture classes used in our experiments.

The previous method demonstrated the adequacy of our image descriptors in simple texture classification tasks. Here we go further and introduce a method that performs classification and segmentation simultaneously. A generative model describes the distribution of the affine-invariant descriptors, along with co-occurrence statistics for nearby patches [17]. The EM algorithm is used to learn the generative model. This allows unsegmented multi-texture images to be incorporated into the training set. At recognition time, initial probabilities computed from the generative model are refined using a relaxation step that incorporates co-occurrence statistics learned at modeling time. Excellent results are obtained for images of an indoor scene and pictures of animals, see figure 8. Class labels are assigned automatically to each region. The regions assigned to each class are shown in the corresponding column.

### 6.3.2 Recognition of object classes

The recognition of object classes has to take into account intra-class variability and select discriminative features. Our approach [15, 16] constructs and selects scale-invariant object parts. Scale-invariant local descriptors are first grouped into basic parts. A classifier is then learned for each of these parts, and feature selection is used to determine the most discriminative ones. This approach allows robust part detection, and it is invariant under scale changes (i.e. neither the training images nor the test images have to be normalized in size). The steps of our approach are as follows:

1. Extract scale invariant local regions with the Harris-Laplace detector and the DoG detector.

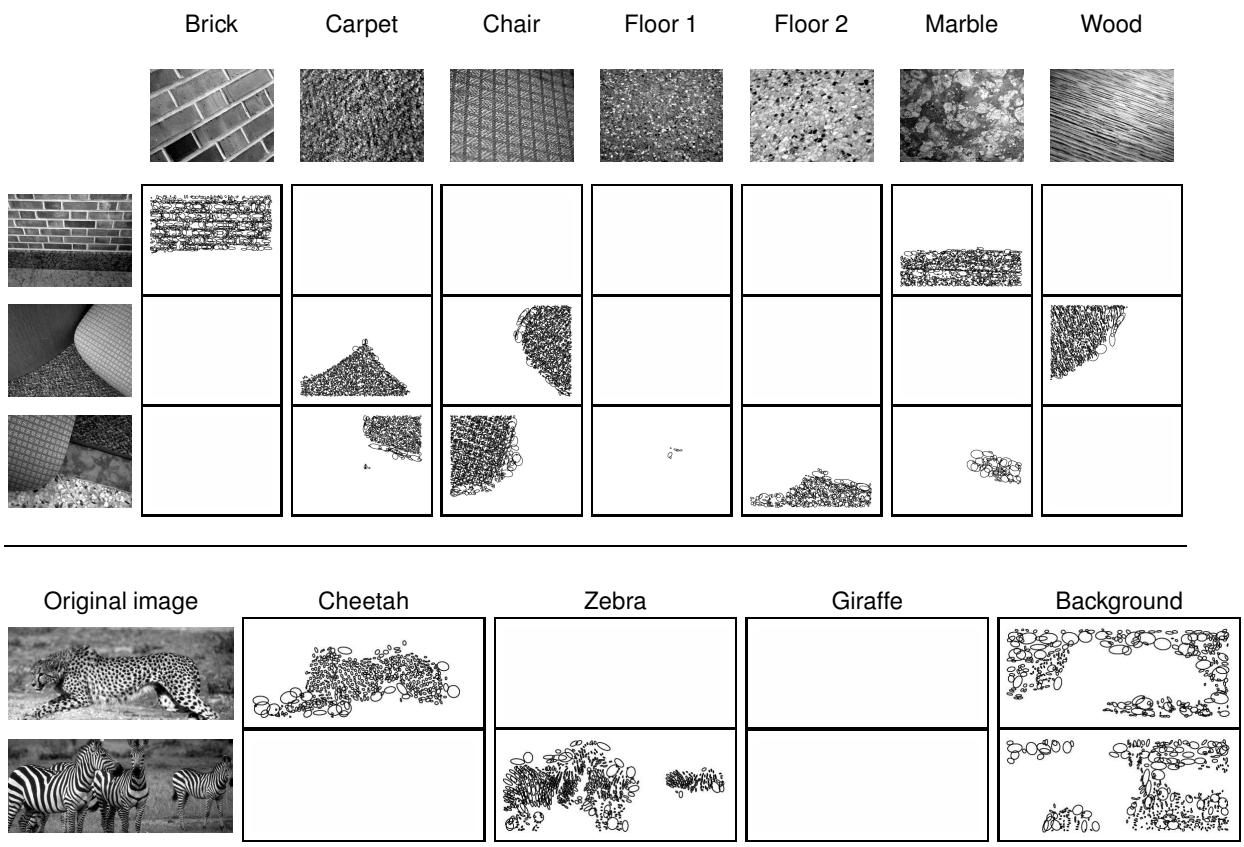


Figure 8: Segmentation/classification results. From top to bottom: a sample image of each texture class from an indoor scene; three successful indoor image segmentation experiments (note that the two marble patches with different orientations are correctly classified); animal image classification examples.

2. Characterize each of these regions with rotation invariant local descriptors; here we use SIFT descriptors.

3. Perform clustering on the positive descriptors to construct a set of object features. In our case (supervised learning) the descriptors of the training images have to be manually sorted into positive and negative ones. Figure 9 shows examples of two different clusters.

4. Learn a part classifier by determining the Gaussian mixture model based on the initial set of clusters.

5. Select features by determining the most distinctive Gaussian components based on likelihood ratio or mutual information. The clusters with the top $n$ scores are selected as significant features for detecting the object class.

The proposed method has been evaluated in a car detection task with significant variations in viewing conditions. Figure 10 shows results on test images (not part of the training set). Note that the

Figure 9: Two different clusters: The first row shows a cluster that represents mainly “tyres”. The second row shows a cluster containing regions detected in the “front window”.

Figure 10: Results on test images using the 10 most discriminant clusters selected by mutual information.

results are very good. Only a few points are incorrectly classified and they could easily be eliminated by any simple coherence criterion. We have also verified that (as in section 6.1.2) SIFT features again give significantly better results than steerable filters.

An extension of this work takes semi-local spatial relations into account and allows affine-invariant part models to be built. The idea is to use graphical models of the characteristic patterns formed by affine-invariant patches to describe salient object parts. Figure 11 illustrates this idea with matches found between face images using the output of the affine-invariant Laplacian and a variant of affine alignment. Note that the patch patterns are stable despite large viewpoint variations and appearance changes.

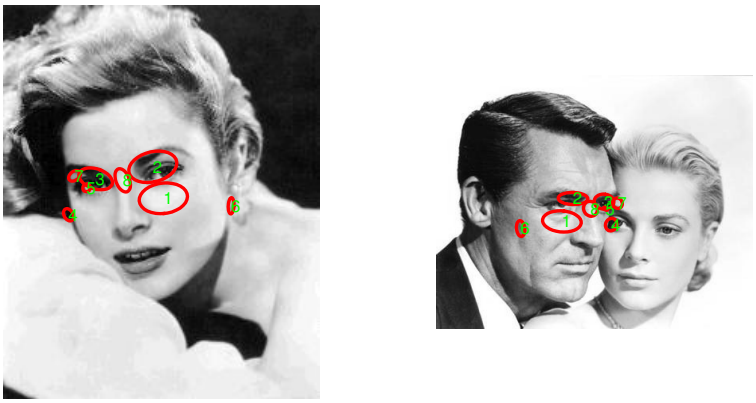

Figure 11: Matching faces with affine-invariant part models.

### 6.3.3 Human detection

Our previous work on human detection [RST02] was based on detecting individual parts and assembling them using dynamic programming. It showed good results in simple conditions, but lacked robustness in general settings. Our new approach models not individual body members, but rather more complex parts that include a larger local context. This increases distinctiveness, while still remaining local enough to allow for occlusion and the detection of close-up views. This is not the case for state-of-art pedestrian detectors based on global information. We also use robust part detectors based on gradient and Laplacian local features that efficiently capture the shape information. Using the probabilistic co-occurrence of these features increases their distinctiveness without reducing robustness. Learning with AdaBoost combines features with the highest co-occurrence probabilities. The detection results are further improved by computing a probabilistic score that takes the relative positions of the parts in the assembly into account. The approach is also very efficient as (i) all part detectors use the same initial features, (ii) a coarse-to-fine cascade approach is used for part detection, (iii) an assembly strategy reduces the number of spurious detections and the search space. The results (see figure 12) are very promising and outperform existing human detectors. Furthermore, the face detection results for frontal and profile views, obtained as one part of the human detector, are comparable with state-of-art detectors [4].

### 6.3.4 Human tracking and action recognition

We have continued our previous work on model based 3D articulated human body tracking from monocular image sequences (e.g. [10]), and opened two new lines of research, on learning dynamical models for 2D (image only) articulated body tracking and on learning to recover 3D human pose reconstructions directly from image silhouettes.

**Model based monocular 3D human pose tracking:** Our past work in this area made it clear that discrete pose reconstruction ambiguities were a far greater source of tracking failure than had previously

[RST02] R. RONFARD, C. SCHMID, B. TRIGGS, "Learning to Parse Pictures of People", *in : Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark*, *IV*, p. 700–714, 2002.

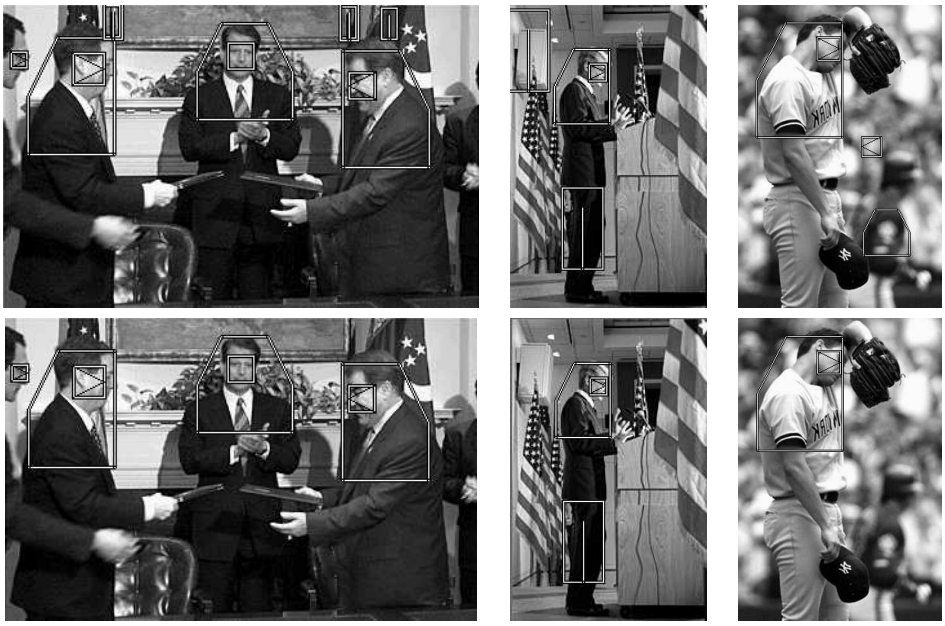

Figure 12: Results for human detection. Top row: body part detection. Bottom row: detection with the joint likelihood model. Note that the joint likelihood model significantly improves the detection results.

been realized. In fact, for any attainable set of image positions of the principal body joints, there are typically some thousands of possible 3D body poses consistent with it[2]. Choosing the wrong initial pose rapidly leads to mistracking, and as the possible solutions continually merge and re-separate in a very complex manner during tracking, it is hard to avoid mistracking. To counter this, we developed [24] a family of multiple hypothesis trackers that use kinematic "flips" to explicitly search the tree of possible poses starting from any given one, either by explicit enumeration, or within a stochastic 'jump-diffusion' style of search processes. When combined with more traditional tracking strategies to handle image correspondence ambiguities, this gives an exceptionally promising human tracker. Some results are shown in fig. 13. The ideas developed in this work are currently undergoing industrial testing (see §7.1).

**Learning dynamical models for 2D articulated human body tracking:** Another strategy is to avoid 3D ambiguities by (initially) tracking only *2D* articulated image motion. The price that one pays is the loss of all 3D-based constraints — most notably rigidity and viewpoint invariance — which leads to its own set of ambiguities. To strengthen the model enough to permit reliable tracking, it is useful to include priors characterizing "typical" 2D human pose and dynamics, and in practice these have quite complex forms so they must be learned from training images. Our work this year [25] has focused on dynamics, as it is this that seems to be needed to track 2D models through changes of aspect (e.g. turns passing from left-side to frontal to right-side views). Our approach learns piecewise linear autoregressive models data by self-consistent clustering of a set of marked-up training sequences.

[2]Basically, each body and limb segment can slant either forwards (towards the camera) or backwards from its root joint. A realistic model has at least 10 segments, and hence at least $2^{10}$=1024 solutions.

Figure 13: Tracking results for a 4 second dance sequence, using our kinematic jump based monocular human tracker. *First row:* original images. *Middle row:* 2D tracking results showing the model-image projection of the best candidate configuration at the given time step. *Bottom row:* the corresponding 3D model configuration rendered from above. Note the good model image overlap and the realistic quality of the reconstructed 3D poses.

Body parameters (34D vectors of 2D joint angles and body lengths) are taken from the training data, reduced to 5–8D (currently by linear PCA) to stabilize later estimation steps, and partitioned using K-means. A (typically degree 1–2) linear autoregressive dynamical is learned for each cluster and lifted to give 34D predictions. Then the data is reclustered according to regularized model prediction accuracy, the models are re-learned, and the whole process is repeated to convergence in an EM-like loop. The overall result is a flexible piecewise model that is capable of capturing the particularities of human dynamics, even for relatively complex motions such as turns. Figs. 14 and 15 show two examples. We are also investigating the use of these kinds of models for classifying different categories of human behaviour.

**Learning to reconstruct 3D human pose from silhouettes:** The above approaches to recovering human pose and movement involve fitting fairly complicated articulated models, which are difficult to initialize and (as we have seen) subject to numerous ambiguities. Perhaps their main default is that they model all *possible* human poses: they do not incorporate much prior knowledge about the much smaller set of poses that are actually *typical*. Hence, they waste a considerable amount of effort searching over poses that are simply (to humans) implausible, and when they mistrack it is often because they followed such a hypothesis too far. Another strategy, which sidesteps both the initialization and the implausibility problems, is to discard the explicit manually-constructed articulated body model, and instead directly *learn* to estimate pose from a suitable set of low-level descriptors extracted from the

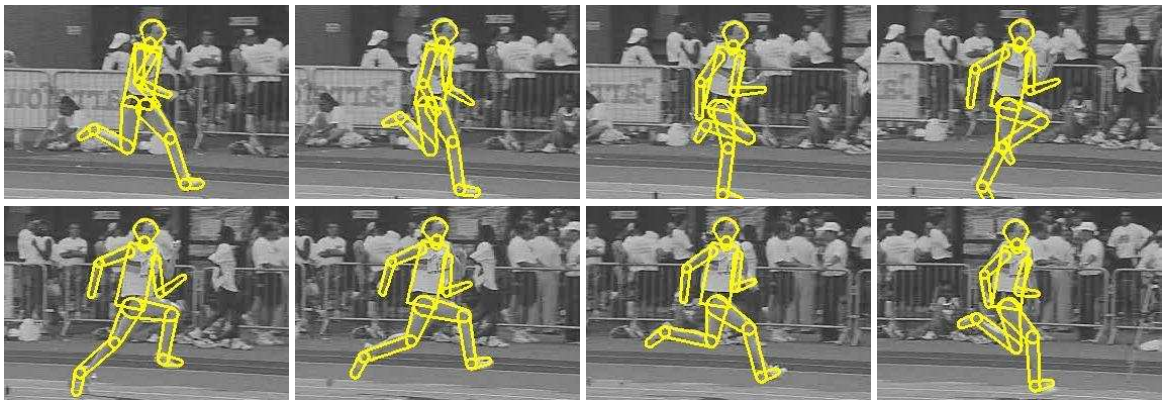

Figure 14: 2D articulated tracking of a running athlete using our learned piecewise autoregressive dynamical model. The tracker was trained using hand-marked images of a different athlete performing a similar motion. The strength of the dynamical model allows individual limbs to be tracked despite the cluttered background, although here the left arm is not tracked very accurately as it is occluded by the body during the initialization phase.

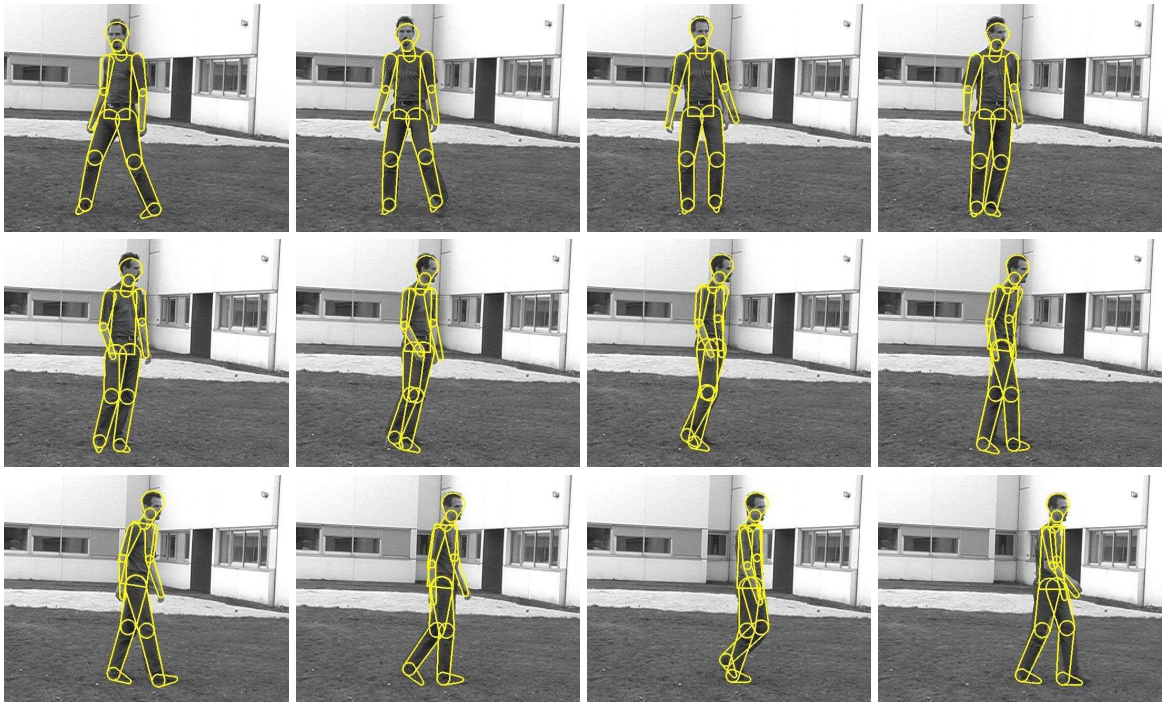

Figure 15: Our 2D articulated tracker running on a walk, turn, and walk sequence. The learned dynamical model smoothly follows the transitions between turning and walking motions.

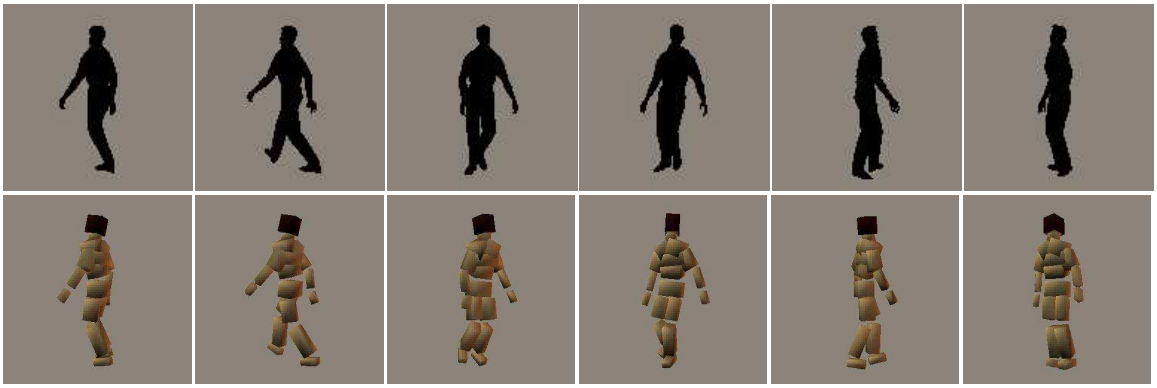

Figure 16: Some sample pose reconstructions from our learning based silhouette to pose regressor, for a spiral walking sequence not included in the training data.

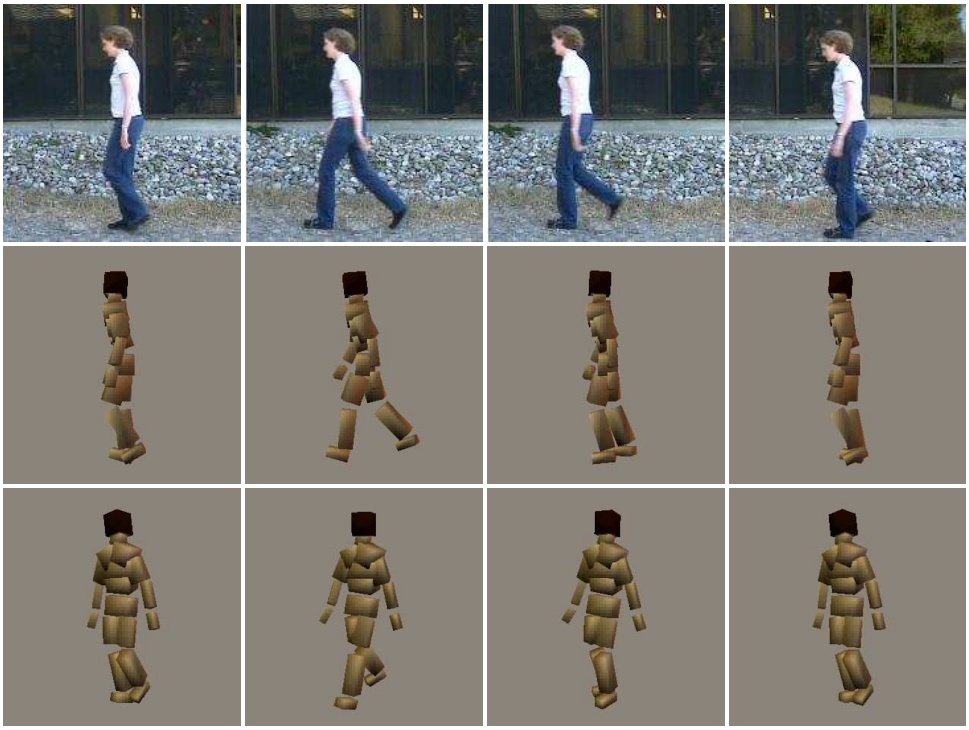

Figure 17: 3D poses reconstructed from some real images (part of a test sequence from `www.nada.kth.se/~hedvig/data.html`). The middle and lower rows show the pose estimates from the original viewpoint and from a new one. The last two columns illustrate limitations of the current implementation. In the third column, a noisy silhouette causes slight mis-estimation of the lower right leg, while the final column demonstrates a case of left-right ambiguity in the silhouette.

image. In our work-in-progress on this, we directly regress 3D full body pose (encoded by joint angle variables) against robust descriptors (histograms of shape-context responses) characterizing the geometry of the human's image silhouette. We have tested various different regression methods. Linear regression (on our very nonlinear descriptors) actually works quite well, but our best results to date are with a sparse Bayesian kernel based method, Relevance Vector Regression. The training data for the regressors is human motion capture data for a variety of human motions, together with the corresponding image silhouettes. The use of realistic training motions ensures typicality of pose, but to eke out our limited quantity of training data, we actually re-render the silhouette images synthetically, to allow training across a wide variety of different viewpoints. The method has been tested on both synthetic images of unseen sequences (to allow ground-truthing) and real images. It achieves average 3D joint angle estimation errors of around 6–7° — significantly better than other existing learning-based approaches, but not yet satisfying from an applications point of view, especially as some of the most important body angles are among the worst estimated. However these are preliminary results and we expect to improve on them significantly. Some example results are shown in figs. 16 and 17.

# 7 Contracts and Grants involving industry

## 7.1 Pandora Studio

**Contributed by**: Bill Triggs, Marius Malciu.

In June 2003 we began an industrial R&D collaboration with the French audiovisual and animation consultancy Pandora Studio, based in Paris and Toulouse. The object is to develop software for semi-automatic 3D human motion capture from single image streams, for production studio applications (film, TV, games). The starting point for the work was the 2002 PhD thesis of Cristian Sminchisescu [Smi02] (supervised by Bill Triggs, and supported by an Eiffel scholarship and the EU project VIBES). Marius Malciu, the Pandora engineer responsible for developing the prototype, visited LEAR in June-July 2003 to begin the work.

## 7.2 Bertin Technologies

**Contributed by**: Frederic Jurie, Cordelia Schmid, Roger Mohr.

In November 2003 we started a new industrial collaboration with the French company *Bertin Technologies*. The object is to develop algorithms for detecting and recognizing objects in the context of an unmanned camera. Applications are typically outdoors military surveillance applications in which hidden cameras have been left to detect the presence of military vehicles. A PhD thesis (CIFRE funds) will start on this subject in the beginning of 2004.

[Smi02] C. SMINCHISESCU, *Estimation Algorithms for Ambiguous Visual Models — Three Dimensional Human Modeling and Motion Reconstruction in Monocular Video Sequences*, PdD Thesis, INPG, July 2002.

# 8 Other grants

## 8.1 National grants

### 8.1.1 Ministry grant MoViStaR

**Contributed by**: Cordelia Schmid, Charles Bouveyron, Jianguo Zhang.

MOVISTAR is a joint national project ("action concertée incitative") under the program "Masses de Données" (Large Quantites of Data). The partners are the INRIA/Gravir project LEAR (C. Schmid), the INRIA project MISTIS (F. Forbes), the SMS team of the LMC laboratory (S. Girard) and the Heudiasyc laboratory (C. Ambroise). MOVISTAR started in September 2003. It aims to develop techniques for mining visual information from large image collections, to achieve reliable recognition of object categories. In particular, it will concentrate on applying and adapting statistical data reduction techniques and on the integration of spatial information.

## 8.2 European projects

### 8.2.1 Vibes

**Contributed by**: Cordelia Schmid, Bill Triggs, Ankur Agarwal, Salil Jain, Michael Sdika, Krystian Mikolajczyk [Oxford].

VIBES (Video Browsing, Exploration and Structuring) is a 5th framework FET-Open project. It started in November 2000 and will run for 3 years, with a 6 month extension. Its main goal is the automatic extraction of high-level representations from video streams. The partners are KTH Stockholm, Sweden (coordinator), LEAR (France), Oxford University (UK), the Katholieke Universiteit Leuven (Belgium), the Ecole Polytechnique Fédérale de Lausanne (Switzerland), and the Weizmann Institute of Science (Israel). VIBES represents an effort of 32 person-years with a total budget of 2.3 MEu, of which 1.7 MEu is funded by the European community. LEAR works on the video indexing, tracking and human reconstruction themes of VIBES.

### 8.2.2 LAVA

**Contributed by**: Cordelia Schmid, Bill Triggs, Roger Mohr, Guillaume Bouchard, Gyuri Dorko, Peter Carbonetto, Michael Sdika.

Learning for Adaptable Visual Assistants (LAVA) is a 5th framework RTD project started in May 2002 for 3 years. It aims to develop advanced machine learning based computer vision techniques for understanding everyday scenes, in particular for applications suited for embedding in camera-equipped electronic devices such as personal assistants and portable telephones. LAVA is an interdisciplinary project, involving teams working on machine learning, computer vision, and cognitive modeling and data fusion. The coordinator is Xerox Research Centre Europe (XRCE, Grenoble, France), and the other partners are: LEAR; VISTA (IRISA-INRIA, Rennes, France); Royal Holloway College and the University of Southampton (Egham and Southampton, U.K.); Lund University (Lund, Sweden); Graz Technical University and the University of Leoben (Graz and Leoben, Austria); the Institut Dalle Molle

d'Intelligence Artificielle Perceptive (IDIAP, Martigny, Switzerland); and the Australian National University (ANU, Canberra, Australia). In total LAVA will involve 51 person-years of research effort, for a total budget of 4.3 MEu, including 2.4 MEu of European Union support. LEAR is working mainly on the development of image descriptors for individual images, the interface between vision and learning, and semi-supervised learning.

#### 8.2.3 PASCAL

**Contributed by**: Bill Triggs, Cordelia Schmid, Gyuri Dorko, Guillaume Bouchard.

PASCAL (Pattern Analysis, Statistical Modelling and Computational Learning) is a Network of Excellence that will start in December 2003 for four years, funded by the Multimodal Interfaces theme of the EU 6th framework. The focus is on applying advanced machine learning and statistical pattern recognition techniques to the analysis of various types of sensed data. It will unite around 120 European researchers in machine learning, pattern recognition, and application domains including computer vision, natural language processing including speech, text and web analysis, information extraction, haptics and brain computer interfaces. The coordinator is Prof. John Shawe-Taylor of Southampton University. Bill Triggs or LEAR is coordinating the computer vision aspects and various other management activities. LEAR and Xerox Research Europe (XRCE) together form one of PASCAL's 14 key sites, focusing on computer vision and language processing.

#### 8.2.4 AceMedia

**Contributed by**: Cordelia Schmid, Navneet Dalal, Bill Triggs.

AceMedia is a 6th framework Integrated Project that will run for 4 years starting from January 2004. Its goal is to integrate knowledge, semantics and content for user-centred intelligent media services. The partners are: Motorola Ltd UK (coordinator); Philips Electronics Netherlands; Thomson France; Queen Mary College, University of London; Fraunhofer FIT; Universidad Autonóma de Madrid; Fratelli Alinari; Telefónica Investigación y Desarrollo; the Informatics and Telematics Institute, Dublin City University; INRIA (including the TexMex team at IRISA in Rennes, Imedia at Rocquencourt in Paris, and LEAR in Grenoble); France Télécom; Belgavox; the University of Karlsruhe; Motorola SAS France.

### 8.3 Bilateral relationship

#### 8.3.1 University of Oxford, UK

**Contributed by**: Cordelia Schmid, Bill Triggs, Krystian Mikolajczyk [Oxford], Andrew Zisserman [Oxford].

The collaboration with the research group of A. Zisserman, university of Oxford, is partially funded by the European project VIBES and by the INRIA postdoctoral fellowship of K. Mikolajczyk. Collaboration focuses include invariant local feature detectors and human detection. In 2003, C. Schmid visited Oxford for a week and K. Mikolajczyk visited Grenoble for a week.

### 8.3.2 University of Illinois at Urbana-Champaign, USA

**Contributed by**: Cordelia Schmid, Gyuri Dorko, Svetlana Lazebnik [UIUC], Jean Ponce [UIUC], Fred Rothganger [UIUC].

The research project on 3D object recognition between the research group of J. Ponce and LEAR is funded by the CNRS/UIUC collaboration. In 2003, S. Lazebnik visited Grenoble twice, for a week each time, and C. Schmid and J. Ponce each visited the partner institution once.

### 8.3.3 Australian National University (ANU), Canberra and National ICT Australia (NICTA)

**Contributed by**: Bill Triggs, Richard Hartley [ANU], Alex Smola [ANU].

This collaboration currently centres around the Australian-funded section of the EU project LAVA, whose focuses are visual methods for recognizing particular locations and kernel based methods for visual recognition. There were no visits between the sites in 2003, but Bill Triggs is planning to visit ANU in spring 2004.

# 9 Dissemination

## 9.1 Leadership within scientific community

- Organization of conference of workshops:
  - General chair of the 9th IEEE International Conference on Computer Vision 2003 (B. Triggs) [1]
  - Co-Organizer of the International Workshop on "Designing Tomorrow's Category-Level 3D Object Recognition Systems", Taomina, Sicily, Italy, September 2003 (C. Schmid)
- Editorial board:
  - International Journal of Computer Vision (C. Schmid)
  - IEEE Transactions on Pattern Analysis and Machine Intelligence (C. Schmid)
  - Machine Vision and Applications (R. Mohr)
  - Techniques et Sciences Informatiques (F. Jurie)
- Area chair :
  - CVPR'03 (B. Triggs)
  - ICCV'03 (C. Schmid)
  - CVPR'04 (C. Schmid)
  - ECCV'04 (C. Schmid)
  - RFIA'04 (C. Schmid)
- Program committee :

  - CVPR'03 (C. Schmid)
  - CAIP'03 (C. Schmid)
  - IJCAI'03 (C. Schmid)
  - TAIMA'03 (C. Schmid)
  - ECCV'04 (F. Jurie and B. Triggs)
  - CVPR'04 (F. Jurie and B. Triggs)
  - RFIA'04 (F. Jurie)

- Other
  - C. Schmid is member of the "INRIA commission d'evaluation"
  - R. Mohr is head of AFRIF (French section of IAPR)

## 9.2 Teaching

- Matching and Recognition, DEA IVR, INPG, 12 h (C. Schmid)
- Multi-media database, 3rd year ENSIMAG, INPG, 11 h (F. Jurie)

## 9.3 Invited presentations

- C. Schmid. *Object recognition based on local invariant descriptors.* The 4th Sino-Franco Workshop on Web Technologies, Taipei, Taiwan, March 2003.
- C. Schmid. *Object recognition based on local invariant descriptors*. Presentation at Toyota, Tokyo, Japan, May 2003.
- C. Schmid and D. Lowe. *Short-course on "Recognition and matching based on local invariant features"*. IEEE Conference on Computer Vision and Pattern Recognition, June 2003.
- C. Schmid. *Object Recognition Based on Local Descriptors*. Workshop on Computational Vision, Rosenon, Sweden, July 2003.
- C. Schmid. *Affine-invariant local features and applications to recognition*. International Workshop on "Designing Tomorrow's Category-Level 3D Object Recognition Systems", Taomina, Sicily, Italy, September 2003.
- R. Mohr. *Visual Recognition: Finding the Needle in the Haystack*. University of Pennsylvenia, Grasp Lab Celebrating a Tribute to Ruzena Bajcsy, 4 October 2003.
- R. Mohr. *Visual Learning, the next Decade Challenge*. Dagstuhl workshop on Cognitive Vision Systems, 26-3 October 2003.
- Demonstration on *Retrieving similar scenes from a video*. Fête de la Sciene, Grenoble, October 2003.

# 10 Bibliography

## Books and Leaflets

[1] K. IKEUCHI, O. FAUGERAS, J. MALIK, B. TRIGGS, A. ZISSERMAN (editors), *Ninth IEEE International Conference on Computer Vision*, Nice, France, IEEE Computer Society Press, October 2003.

## Articles and Book Chapters

[2] A. BARTOLI, N. DALAL, R. HORAUD, "Motion Panoramas", *Journal of Visualization and Computer Animation*, 2003, to appear.

[3] G. BOUCHARD, S. GIRARD, A. IOUDITSKI, A. NAZIN, "Nonparametric estimation of the support boundary through linear programming", *Automation and Remote Control*, 2003, to appear.

[4] R. CHOUDHURY, C. SCHMID, K. MIKOLAJCZYK, "Face detection and tracking in a video by propagating detection probabilities", IEEE *Transactions on Pattern Analysis and Machine Intelligence 25*, 10, 2003, p. 1215–1228.

[5] Y. DUFOURNAUD, C. SCHMID, R. HORAUD, "Image matching with scale adjustment", *Computer Vision and Image Understanding*, 2003, to appear.

[6] P. GROS, C. SCHMID, "La reconnaissance des formes dans les images", *in : Perception visuelle par imagerie vidéo*, M. Dhome (editor), Hermès, 2003.

[7] K. MIKOLAJCZYK, C. SCHMID, "Scale and affine invariant interest point detectors", *International Journal of Computer Vision*, 2003, accepted.

[8] C. SCHMID, "Recognition with local photometric invariants", *in : Trends and Advances in Content-Based Image and Video Retrieval*, L. Shapiro, H. Kriegel, and R. Veltkamp (editors), Springer, 2003, to appear.

[9] C. SCHMID, "Weakly supervised learning of visual models and its application to content-based retrieval", *International Journal of Computer Vision 56*, 1, 2003, p. 7–16.

[10] C. SMINCHISESCU, B. TRIGGS, "Estimating Articulated Human Motion with Covariance Scaled Sampling", *Int. J. Robotics Research 22*, 6, June 2003, p. 371–391, Special issue on Visual Analysis of Human Movement.

## Conference and Workshop Publications, etc.

[11] G. BOUCHARD, G. CELEUX, "Supervised Classification with Spherical Gaussian Mixtures", *in : CLADAG meeting*, Italian Statistical Society, september 2003. Invited paper session.

[12] G. BOUCHARD, "Localised Mixtures of Experts for Mixture of Regressions", *in : Between Data Science and Applied Data Analysis*, Schader, Gaul, Vichi (editors), Springer-Verlag, p. 155–164, 2003.

[13] P. CARBONETTO, N. DE FREITAS, P. GUSTAFSON, N. THOMPSON, "Bayesian feature weighting for unsupervised learning, with application to object recognition", *in : Proceedings of the Workshop on Artificial Intelligence and Statistics*, 2003.

[14] P. CARBONETTO, N. DE FREITAS, "Why can't José read? The problem of learning semantic associations in a robot environment", *in : Proceedings of the NAACL Human Language Technology Conference Workshop on Learning Word Meaning from Non-Linguistic Data*, 2003.

[15] G. DORKO, C. SCHMID, "Feature selection for object class detection", *in : The Learning Workshop, Snowbird, Utah*, 2003.

[16] G. DORKO, C. SCHMID, "Selection of Scale-Invariant Parts for Object Class Recognition", *in : Proceedings of the 9th International Conference on Computer Vision, Nice, France*, *1*, p. 634–640, 2003.

[17] S. LAZEBNIK, C. SCHMID, J. PONCE, "Affine-Invariant Local Descriptors and Neighborhood Statistics for Texture Recognition", *in : Proceedings of the 9th International Conference on Computer Vision, Nice, France*, *1*, p. 649–655, 2003.

[18] S. LAZEBNIK, C. SCHMID, J. PONCE, "Representing, learning, and recognizing non-rigid textures and texture categories", *in : The Learning Workshop, Snowbird, Utah*, 2003.

[19] S. LAZEBNIK, C. SCHMID, J. PONCE, "Sparse Texture Representation Using Affine-Invariant Neighborhoods", *in : Proceedings of the Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, USA*, *2*, p. 319–324, 2003.

[20] K. MIKOLAJCZYK, C. SCHMID, "A performance evaluation of local descriptors", *in : Proceedings of the Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, USA*, *2*, p. 257–263, June 2003.

[21] K. MIKOLAJCZYK, A. ZISSERMAN, C. SCHMID, "Shape recognition with edge-based features", *in : Proceedings of the 14th British Machine Vision Conference, Norwich, England*, *2*, p. 779–788, September 2003.

[22] J. PONCE, F. ROTHGANGER, S. LAZEBNIK, K. MCHENRY, C. SCHMID, S. MAHAMUD, M. HEBERT, "3D Photography from Photographs and Video Clips", *in : International Symposium on Core Research for Evolutional Science, Technology (CREST) — Ikeuchi Project, Tokyo, Japan*, p. 153–182, 2003.

[23] F. ROTHGANGER, S. LAZEBNIK, C. SCHMID, J. PONCE, "3D Object Modeling and Recognition Using Affine-Invariant Patches and Multi-View Spatial Constraints", *in : Proceedings of the Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, USA*, *2*, p. 272–277, 2003.

[24] C. SMINCHISESCU, B. TRIGGS, "Kinematic Jump Processes For Monocular 3D Human Tracking", *in : IEEE Int. Conf. Computer Vision and Pattern Recognition*, June 2003.

## Miscellaneous

[25] A. AGARWAL, *Learning Dynamical Models for Tracking Complex Motion*, Mémoire, Institut National Polytechnique de Grenoble, June 2003.